UNIVERSITY OF WASHINGTON

PRP-31N

An Experimental Critique of the Method of

Constant Stimuli and Some Alternative Procedures

by

Don A. Ronken

University of Washington

This report has been prepared under contract NONR 477(34) between the Office of Naval Research and the University of Washington. Inquiries concerning the contents of this report should be made to the Principal Investigator, Eugene Galanter, University of Washington, Seattle, Washington 98105.

PRP-31N

An Experimental Critique of the Method of Constant Stimuli and Some Alternative Procedures[1]

Don A. Ronken

University of Washington

It is often said that different psychophysical procedures produce quantitatively different results. In fact, it is considered an accomplishment for a theory to be able to predict the results of one experiment from another which uses precisely the same signals (Luce, 1963). Most often, however, methodological work as such is placed on the opposite end of a value scale from what are called substantive problems. For many practical experimenters concerned about getting on with work involving more substantive issues, the choice of psychophysical method is not of overwhelming concern.

Why should there be any misgivings about traditional methods which have been used for more than a century? The answer is found in the presumed validity (not yet established) of the following assertion: the influence of the psychophysical procedure becomes more important as the effect of the independent variable becomes more and more difficult to detect. Large effects might be expected to manifest themselves in spite of unfavorable conditions, but such [illegible] not the case for elusive phenomena like the classical "constant error" or "[illegible]me error." This report is an experimental history of a search for such small effects and some difficulties encountered with the traditional procedures.

After a series of experiments it became evident that the assumptions usually made about the experimental procedures for discrimination were not being met in practice. The breakdown of these assumptions should not be taken lightly, for

they are the conditions which form the axioms of theories for discrimination (Luce and Galanter, 1963). In addition, these experiments suggest that the observer plays a more active role than had been thought, which introduces considerable complexity into behavior obtained using the classical methods.

The assumptions often made about discrimination are evident from the procedures of the method of constant stimuli, the traditional experimental design in which two signals are presented as an ordered pair and the subject is required to produce some sort of comparative response. Trials are almost invariably considered to be independent. The variables thought to be relevant under these conditions are the physical relationship between the two signals and what may be called separation variables, such as the distance or time between signals, interpolated stimuli and so forth. We limit our investigation to the case of temporal separation. These conditions lead us to expect that the only stimuli relevant to a response on any given trial are those the experimenter presents on that trial, and that the length of the temporal interval between the paired signals is a significant variable. The first two experiments to be described here failed to confirm these basic assumptions of the method of constant stimuli.

It will be recalled that the constant error represents the failure of the subject to identify two objectively equal stimuli as being psychologically equal. For example, the second presentation of the same tone will be reported to sound louder than the first. Experiments demonstrating these effects lead to much controversy in the 1930's (Woodworth, 1938) which has not yet been resolved, either empirically or theoretically (Luce and Galanter, 1963). This program of research, in contrast to some of the earlier work on the constant error, sought asymptotic behavior from individual subjects. Six pilot experiments are reported

here which were performed over a two year period. No one of these taken singly provides a crucial demonstration of the points to be made, but the entire sequence appears consistent enough to increase the credibility of results from individual experiments.

The original purpose of the first experiment was to evaluate the constant error at several widely spaced intensities of the standard stimulus to determine whether the constant error would follow Weber's Law, an hypothesis proposed by Galanter (1962) and theoretically justified by Luce and Galanter (1963). These data were mentioned in an earlier report (Ronken and Galanter, 1965), but a more complete analysis can be given at this time. This experiment demonstrates the failure of the assumption that the only stimuli relevant to the response are those presented on the trial when the response occurs. As will be seen, the constant error for a stimulus pair depends strongly upon other stimuli as well.

Experiment 1

The method of constant stimuli was used in its barest essential form for experiments 1 and 2. That is, the stimulus presentation on each trial was of the form $(S, S \pm \Delta)$: a standard stimulus followed by one of two comparison stimuli which differed from the standard by an amount Δ . All the stimuli were 1000 Hz tones produced by a Hewlett-Packard 200 AB oscillator which was operated at a fixed level. Different amplitudes were obtained via precision attenuators selected by relays which switched during periods when the signal was of . Miscroswitches, preceding the attenuators, and activated by the cams of a motor-driven timer, turned on each of the tones for one second and provided a one second inter-stimulus interval within each trial. Switching transients were reduced by passing the signal through the narrow bandpass filter section from a

General Radio tuning fork oscillator, model 813-A. The waveform was symmetrical with rise and fall times of about 5 msec. that did not produce any audibly noticeable transient, in agreement with findings reported by Wright (1960). The earphones were two Permoflux PDR-8 receivers wired in series and in phase for binaural presentation and mounted in Willson Sound Barrier earmuffs fitted with fluid-filled cushions.

Signals throughout this report are designated as relative to a comfortable listening level called S_o, which was about 70 db. Sensation Level. The observer was seated in a double-walled IAC sound attenuating booth, a requirement for these experiments because noise was not added to the background. On each trial, a small white pilot lamp gave an unobtrusive warning signal one second before the onset of the first tone; there were no markers indicating signal duration. Following the traditional paradigm, no feedback of correct responses was given. The automatic apparatus cycle of all these experiments was contingent upon occurrence of a response, so the trials were self-paced for the subject.

Instructions to the observer were to depress a pushbutton labeled "Louder" if he thought the second tone of the trial was louder than the first, or the "Softer" button if the second seemed softer.[2] The probability that the second signal of the pair had the larger amplitude was always 1/2. No additional constraint was placed upon construction of the random sequences, which were assembled by a digital computer, using a psuedo-random number generator. The subject was given all details regarding construction of the sequence, and following a few blocks of trials, an attempt was made to explain the "gambler's fallacy." An average of five blocks of 100 trials each, interspersed with rest periods, comprised an experimental session of about ninety

minutes. The subjects in all these experiments were undergraduates whose wages were determined from symmetric payoff matrices, except that in experiment 4, the writer served as the observer. Data from 10 sessions were discarded before the data reported here were collected because analysis of block-to-block variability suggested that the subject had not reached asymptotic performance.

It has been observed in these discrimination experiments that practice effects often are present over many experimental sessions and thousands of trials. These conditions are quite different from reports of the role of practice in detection experiments. Green and Swets (1966) present several references indicating the relatively minor effect of practice on detection data and only one abstract citation (Tanner and Rivette, 1963) to suggest that the case may be different for studies using more complex signals.

Condition 1.

With the standard stimulus set at S_o and Δ at 0.4 db., 500 observations were collected. The half-filled circle of Figure 1 displays the obtained result in the form of a unit square plot. For this figure, the abscissa is the probability of a "Louder" response, given that the second signal had the smaller amplitude, that is, the presentation was $(S, S-\Delta)$. The ordinate is the probability of a "Louder" response given presentation $(S, S+\Delta)$.

Condition 2.

Four standard stimuli were selected at intensities which were -20, -10, 0 and +10 db. re S_o. The same Δ, 0.4 db., was used at each standard. These standards formed 8 possible presentations, $(S_{-20}, S_{-20} \pm \Delta)$; $(S_{-10}, S_{-10} \pm \Delta)$;

$(S_o, S_o \pm \Delta)$; $(S_{+10}, S_{+10} \pm \Delta)$ which were selected at random with probability 1/8 to generate stimulus ensemble A. Data from an experimental session using ensemble A are shown in Figure 1 connected by the solid line, where each point represents 100 trials. The lines connecting points in this figure are only to indicate that such sets of points form a condition of the experiment. No further significance is to be attached to these lines.

Condition 3.

Presentations $(S_{+10}, S_{+10} \pm \Delta)$ were used, as in condition 1, during one session. Likewise, during another session, only presentations $(S_{-10}, S_{-10} \pm \Delta)$ occurred. These data appear as the other two half-filled points in Figure 1 and are based on 100 trials each.

Condition 4.

Stimulus ensemble B was created by selecting four standards around S_o as before, but covering a smaller range of intensities. These standards were -2, 0, +2 and +4 db. re S_o. Figure 1 shows the data points from this ensemble, representing 100 trials each, connected by a dashed line.

Results of experiment 1.

A response bias, indicating the presence of a constant error, would be indicated in Figure 1 by displacement of the data along an isosensitivity curve, away from the minor diagonal connecting the points (0, 1) and (1, 0). Points above the diagonal indicate a preponderance of "Louder" responses and therefore the usual sort of constant error (Woodworth, 1938). Three data points of Figure 1, each one representing discrimination using S_o as the standard stimulus, indicate the effect of other stimuli upon the discrimination of S_o itself. For

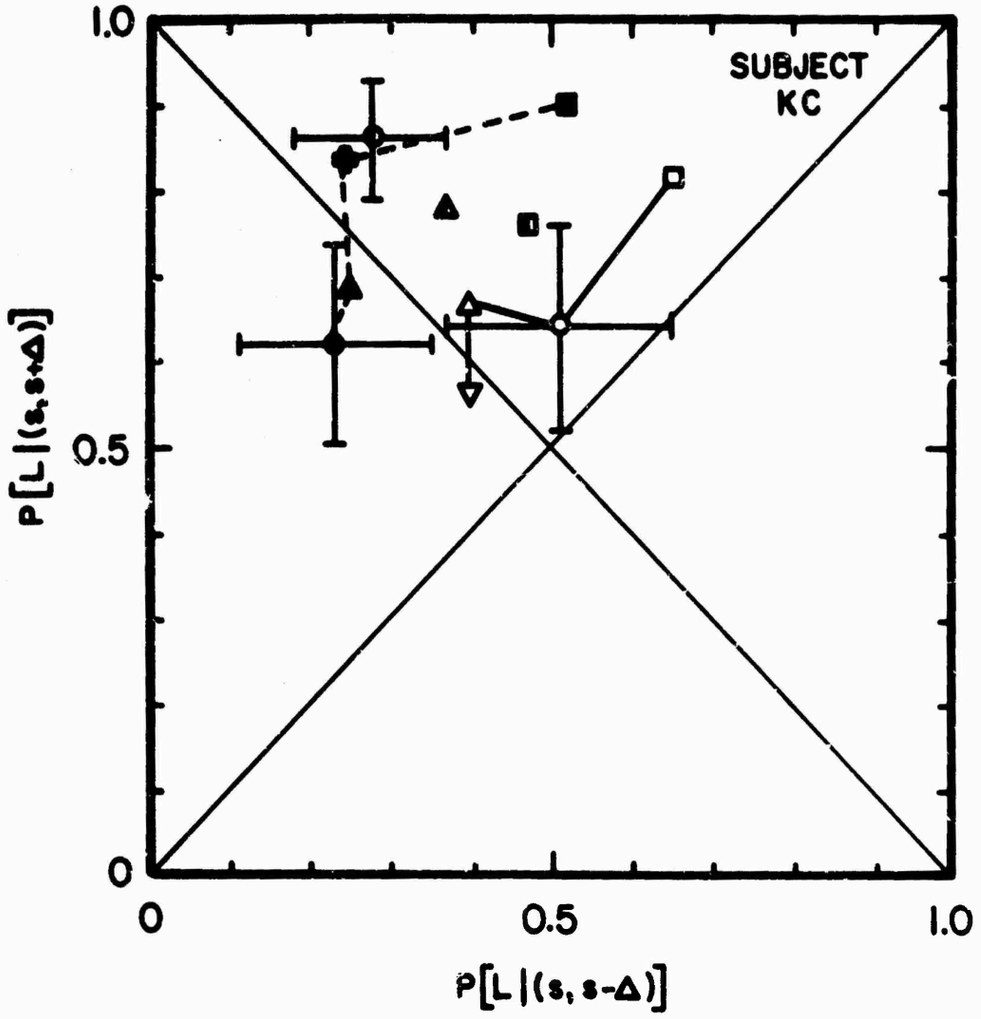
these three points, the 95% confidence limits are also shown. These are the intervals defined by limits ± 2 standard deviations from each experimentally determined proportion. Note that the discrimination of S_o varies widely, both in terms of response bias and in terms of signal strength, as a function of the stimulus ensemble of which it is a member. The same effect occurs for standards S_{-10} and S_{+10}. Condition 4, using the narrow-range ensemble, indicates that the physical difference between standards need not be large to produce these ensemble effects.

Discussion of experiment 1.

The data of Figure 1 indicate the operation of a powerful variable not specified by the assumptions ordinarily made about discrimination. Evidently the effect of a stimulus on any given trial is not independent of stimuli used on other trials in the same experiment. The presence of other pairs makes a given stimulus pair less discriminable and at the same time introduces changes in response bias. From the standpoint of signal detection theory, this is a rather unusual result. In detection experiments, the theory has been very successful in demonstrating that a single experimental variable affects either sensitivity or response bias parameters, but not both. Experiments 3 and 4 will present some additional data on this compound effect of the ensemble.

Experiment 2

The first experiment showed that the relevant physical parameters of discrimination experiments may be more complicated than the method of constant stimuli suggests. Experiment 2, on the other hand, presents some data regarding an assumption often made about a separation variable of the classical method. This assumption is that the temporal interval between stimuli which are to be discriminated is an effective variable. It is obvious that unless these assumptions



Amplitude of standard in db. re S_o	Single Standard	Ensemble A	Ensemble B
	+10 ■	+10 □	+4 ■
	0 ●	0 ○	+2 ⬢
	-10 ▲	-10 △	0 ●
		-20 ▽	-2 ▲

Figure 1: Data from the method of constant stimuli under single standard and ensemble conditions.

of the relevant variables for discrimination are valid, any phenomena which assumes the existence of comparison responding, such as the constant error, cannot be examined in the absence of confounding effects.

The apparatus and procedures of experiment 1 were used with presentations ($S_o, S_o \pm \Delta$) and $\Delta = 0.4$ db. The inter-stimulus interval was fixed for a block of trials, but different blocks used 1, 3, 6 or 12 second intervals in an irregular order. Following 7 practice sessions of about 400 trials per session, the data shown in Figure 2 were collected from the subject of experiment 1. Each of these points is based on 200 trials, except the point for an inter-stimulus interval of 1 second, which represents 500 trials. The 95% confidence limits are shown for the 6 and 12 second conditions.

Discussion of experiment 2.

Intuition suggests that the points of Figure 2 could have come from the same isosensitivity curve. To examine this hypothesis in a more quantitative fashion, we use a significance test recently developed by Gourevitch and Galanter (1966).[3] The results of the test agree with our intuition; for example, it indicates that the probability is greater than 0.5 that the 6 and 12 second points could have been obtained by repeated sampling from the same underlying distributions. The conclusion is drawn that the inter-stimulus interval has not affected the discriminability.

Results similar to those of experiments 1 and 2 can be found in the literature, although asymptotic data for individual observers are almost nonexistent. The potential usefulness of many of these studies is reduced because no measures of discriminability (JND, DL, etc.) are reported (Koester and Schoenfeld, 1946; Needham, 1934, 1935; Postman, 1947). Other, often-cited papers present only

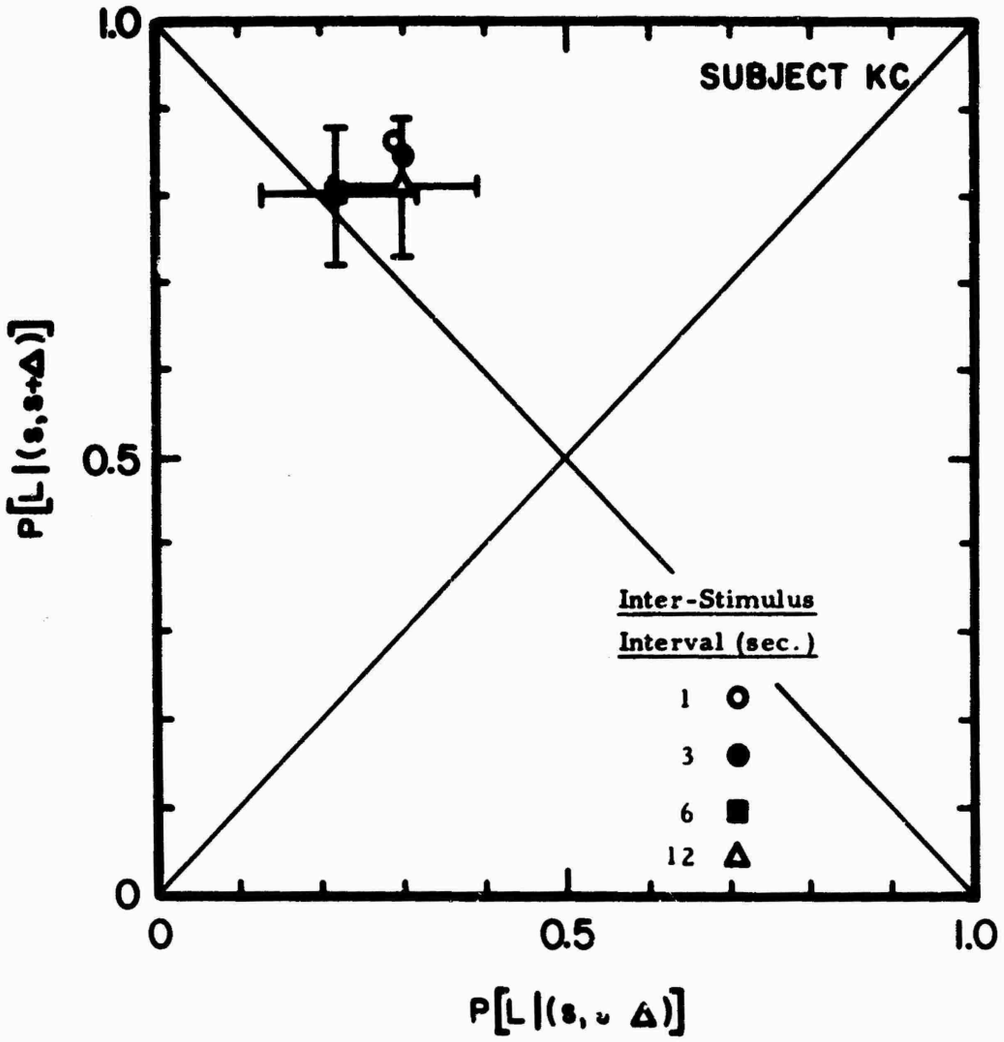


Figure 2: Data from the method of constant stimuli using a single standard at four different inter-stimulus intervals.

results obtained by averaging across subjects (Harris, 1948, 1952a; Postman, 1946, 1947). The fallacy involved in averaging data from different observers can be appreciated by observing the radically different shapes of the empirical functions which are obtained from individual subjects in the constant error type of study (Harris, 1949; Koester and Schoenfeld, 1946; König, 1957; Needham, 1934).

We shall not review here the work which seems to be related to the ensemble effect of experiment 1 (Doughty, 1949; Harris, 1948, 1952a; Harris, Pikler, Hoffman and Ehmer, 1958; Needham, 1935; Pollack, 1954, 1956; Rosenblith and Stevens, 1953; Woodrow, 1933). Similarly, we shall only mention some experiments which have found that increasing the inter-stimulus interval has very little effect on discriminability (Harris, 1949, 1952a; König, 1957; Pollack, 1954; Postman, 1946; Whipple, 1901). Regarding this last point, it might be argued that the time between stimuli has no effect on discriminability because the range of temporal intervals investigated was inappropriate. Against this stands the intuitive feeling that for less than perfect discrimination, time factors on the order of a few seconds ought to be relevant. Judging from the large numbers of studies which have varied the time between stimuli with the intention of observing some effect, many authors have found the second view more compelling.

The shortcomings of the method of constant stimuli, which were shown in the first two experiments and illustrated by examples from the literature, are not obvious so long as the conventional procedure is rigidly followed. Unexpected effects have occurred only in the detailed data of experiments using expanded versions of the method. This result is not interpreted as establishing the useful

bounds of the classical method, but rather as suggesting that there are some serious underlying complications.

The nature of part of the complexity seems to have been suggested by Wever and Zener (1928), who found that observers could produce discrimination-like limens from a procedure which involved presentation of only single stimuli. However, their results had suprisingly little effect on the way experiments were done. Not even the more recent work on identification of single elementary stimuli, much of it within the context of information theory (Garner, 1953; Garner and Hake, 1951; Eriksen and Hake, 1955) has deterred experimenters from using the classical method. This generalization must be qualified to provide exceptions for the prodigious methodological efforts of Harris (1948) and Pollack (1954, 1956).

Harris has recognized some of the problems associated with the method of constant stimuli and has tried many procedural variations on large numbers of subjects in attempting to rescue the technique (Harris, 1948). Unfortunately for our purpose, this work has not used experienced observers nor reported individual data. Harris concludes from his pitch discrimination work that the best procedure is to vary the standard stimulus from trial to trial in a slow, regular fashion over some range in the vicinity of the standard (he used the range from 750 to 850 Hz). It would be informative to examine the data of this procedure in detail to observe whether the response bias tracks the standard, as it did in experiment 1, conditions 2 and 4. Regarding this point, Pollack (1956) has presented some data for amplitude discrimination of a 1000 Hz tone in which the standard varied at random from trial to trial over several ranges. His graphs (Pollack, 1956, figure 2, page 908) for highly practiced observers show

that under such conditions, the response bias depends strongly upon the amplitude of the standard, independently of the amplitude of the comparison, as found in experiment 1.

Harris (1948) has also examined the technique of varying the standard at random, but rejects the procedure in favor of the gradually ascending and descending method on the grounds that the random technique produced greater experimental variability across 50 subjects. In addition to questioning the validity of this criterion for the rejection of the random method, it is possible to take issue with the decision on the basis of some of Harris' own data. Procedures 15 and 16 (Harris, 1948, table 1, page 316) used the method of single stimuli, but selected the signals according to the random or gradually varying procedures. The average DL obtained with the gradually changing standard was 1/2 to 1/3 the size of the random DL, and it was possible for the subject to be correct 75% of the time under the gradual procedure but not under the random method. This result suggests considerably more partial identification behavior is possible when conditions change gradually and in a regular fashion than when changes occur at random. The term partial identification (Bush, Galanter and Luce, 1963) is used to mean that the subject does not identify each stimulus with a unique label but classifies each as being high or low in pitch, loud or soft, etc. The importance of this behavior for discrimination will now be demonstrated.

The role of partial identification behavior.

Let us consider how the ability of human observers to identify single stimuli as being high or low on a continuum might affect results of experiments using the method of constant stimuli. This procedure uses a presentation set of the form

$$\{(S, S \pm \Delta_1);\ (S, S \pm \Delta_2);\ \ldots\ (S, S \pm \Delta_i)\} \quad .$$

That is, every pair of signals is composed of the standard followed by another signal differing from the standard by an amount $\pm\Delta_1$, $\pm\Delta_2$, ... $\pm\Delta_i$, where $\Delta_1 < \Delta_2 < \ldots < \Delta_i$. Suppose the observer chose to disregard the standard entirely and base his response only upon the second signal by locating it on the continuum with a partial identification response. Data from the method of single stimuli indicate that such responses can be made with resolving power not very different from that of paired signal discrimination (Bressler, 1933; Harris, 1948; Pollack, 1954; Wever and Zener, 1928). Moreover, note that the pair which the experimenter expects to be most difficult to discriminate, $(S, S \pm \Delta_1)$, has second signals which are separated by an amount $2\Delta_1$. By responding "Louder" to the presentation which has $(S+\Delta_1)$ as the comparison stimulus and "Softer" to the pair having $(S-\Delta_1)$, the observer could take advantage of the correlation between overall loudness and correct discrimination responses which exists in the method of constant stimuli. In the extreme case in which these second signals are not confused when partially identified, the scheme would lead to 100% correct responding. More realistic cases of imperfect identification will lead to less than perfect responding; how much less is of course an empirical question. We may refer to this hypothesized form of behavior as quasi-discriminative responding.

It might be anticipated that such variables as practice and feedback for correct responses would affect the importance of this response system. Because these quasi-discriminative responses are assumed to be independent of the standard stimulus, they also might be expected to become more influential when actual comparison of first and second signals is made more difficult by such procedures as using a long inter-stimulus interval or interpolating distracting

stimuli between the paired signals.

The results of experiment 1 were particularly suggestive of the presence of partial identification responses. This can be seen from the response bias obtained under ensemble conditions 2 and 4 in Figure 1. Note that the presentations with the more intense standards produce a large percentage of "Louder" responses, independent of the value of the comparison stimulus. The response bias follows the overall loudness of the presentation, rather than the standard-comparison relation. A similar result has been cited in other experiments (Needham, 1935; Pollack, 1956; Woodrow, 1933). Experiment 2, which used only a single standard stimulus, also suggested that the first signal was ineffective and that responding was largely on the basis of the second stimulus alone.

Experiment 3 was thought at first (Ronken and Galanter, 1965) to provide a method which would allow such data to be partitioned into discriminative and quasi-discriminative response effects. The manner in which this was to be done may be described most easily after the experimental arrangement has been specified.

Experiment 3

An improved audio circuit presented two 500 msec. bursts of 1000 Hz which could be separated by an inter-stimulus interval of 1, 4 or 8 seconds. The timing functions were performed by solid state timers (Saslow and Markowitz, 1964) which were calibrated by a Hewlett-Packard 523 CR electronic counter. Repeated checks on the accuracy of these timers have shown that the actual range of times produced is less than 1/2% of the mean setting for times greater than 500 msec. The signal switching was performed at sine-zero crossings by an electronic switch that provided an on-to-off ratio of approximately 80 db. Symmetric rise and fall times were obtained by using the narrow bandpass filter.

Stimulus control was improved over the earlier experiments by using an oscillator with increased amplitude stability, Optimation model RCD-2R, which was isolated from the switch by an amplifier and matching transformer. After the filter the signal was again amplified before passing to a bank of 8 T-pad attenuators, each of which provided an independent setting. A fixed amount of attenuation was added before the two PDR-8 receivers, wired in series and in phase.

To calibrate the T-pads, a resistive mixing network was inserted into the circuit in the position ordinarily occupied by the T-pads. The two outputs of this network provided identical inputs for the T-pad circuit and for a precision, 0.1 db. step Daven attenuator. The outputs of the Daven and the T-pads were compared by connecting them to the differential preamplifier of a Tektronix oscilloscope. The horizontal sweep was driven from the signal input to the mixing network. A Lissajous figure served as the null indicator and permitted deviations of less than 0.05 db. to be easily detected (method suggested by M.G. Saslow).

A small pilot lamp provided a warning signal to the subject; the light was on for 1/2 second and went off 1 second before the first tone. The observer's responses on "Louder-Softer" pushbuttons were immediately followed by illumination of pilot lamps which indicated the correct response. Stimulus and response events of each trial were recorded on a digital printer.

The stimulus sequences were obtained from perforated paper tape, 110 trials per block. Each block contained 100 trials for which $\Delta = 0.4$ db. and 10 practice trials with Δ set at 3.0 db. Figure 3 illustrates how the T-pad attenuators were used to form the stimulus presentations. The abscissa

represents the value of the first signal and the ordinate the value of the second, both are decibel scales of the voltage applied to the earphones. Signal levels 2, 3, 4 and 5 were equally spaced in decibels. The setting of T-pad 3 corresponded to S_o of experiments 1 and 2. Each point of Figure 3 indicates a particular stimulus presentation: (2, 3) indicates that T-pad 2 was used for attenuation of the first signal and T-pad 3 for the second. The diagonal line partitions the presentations into those for which the correct response is "second tone louder" (presentations above the diagonal) and those for which the correct response is "softer" (below diagonal). Practice trials, not included in the later analysis, were presentations (3, 6) and (6, 3).

The critical point of this experimental arrangement is that the six presentations of Figure 3 stand in a particular relationship to one another. Consider the case of a highly practiced observer in the following "thought" experiment, which will be recognized as an application of the partial identification behavior referred to earlier. The only presentations in this "thought" experiment are (3, 4) and (3, 2), which occur at random with equal probability. On each trial, the tones are separated by a long inter-stimulus interval, so that the task is difficult, even for our highly practiced subject. Note that this paradigm is the abbreviated version of the method of constant stimuli, for the first signal is always the same and the difference between the second signals is twice the difference between the standard and comparison tones. Obviously the conditions are conducive to what we have called quasi-discriminative responses.

An identical "thought" experiment could be performed by using the presentations connected by the other vertical line in Figure 3, with the same outcome. On the other hand, examination of the stimulus ensembles connected by horizontal lines reveals quite a different situation. For instance, in the case of

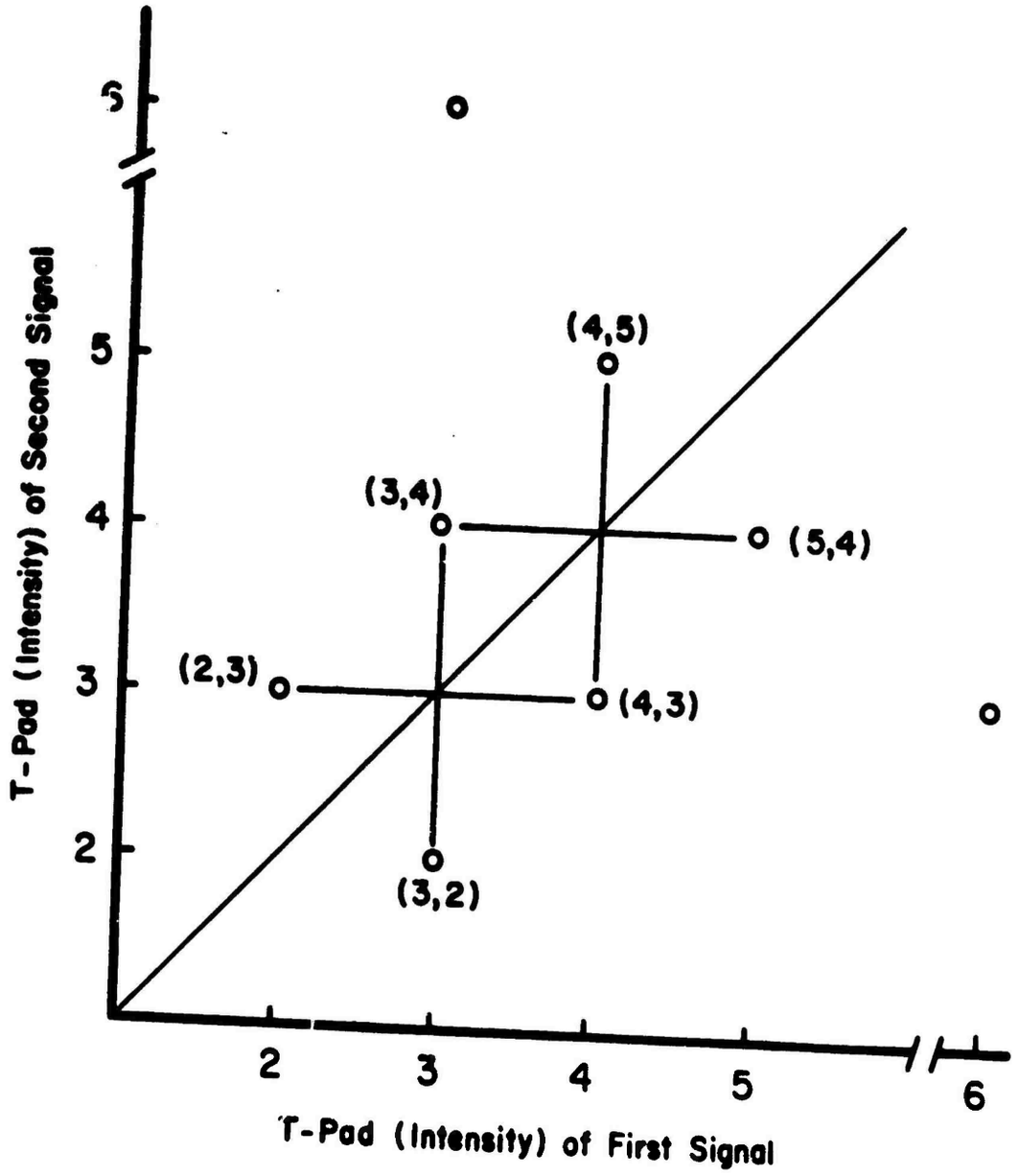


Figure 3: Stimulus presentation scheme for experiment 3.

(2, 3) and (4, 3) , correct identifications of the second signal will not provide any information about which response should be made.

Consider now the six presentations of Figure 3 combined to make a single ensemble, with each presentation occurring with equal probability. The four presentations connected by horizontal lines will not be subject to identification schemes in the same sense as those connected by vertical lines. On this basis, (3, 2) and (4, 5) ought to be more subject to quasi-discriminative responses than the other presentations because these two do not appear in any horizontal pattern. On the assumption that conditions which permit partial identification to lead to correct responses can only facilitate the combination behavior, the prediction is made that correct responses to (3, 2) and (4, 5) will be more frequent than to the other four presentations. In addition, the superiority of (3, 2) and (4, 5) should increase as the discrimination becomes more difficult, as, for example, when the inter-stimulus interval is increased. This last statement comes from the fact that the partial identification responses would be independent of the inter-stimulus interval. The accuracy of these conjectures can be evaluated by examination of the proportion of correct responses for the six presentations individually.

A different observer practiced for 22 sessions, making some 6700 responses to the presentations of Figure 3 at inter-stimulus intervals of 1, 4 and 8 seconds before the final data were collected. The inter-stimulus interval was the same throughout a block of trials, but varied over blocks and sessions at random.

Results of experiment 3.

Figure 4 presents the proportion of correct responses, P(c), for each of the six presentations, averaged over sessions. The solid points are data

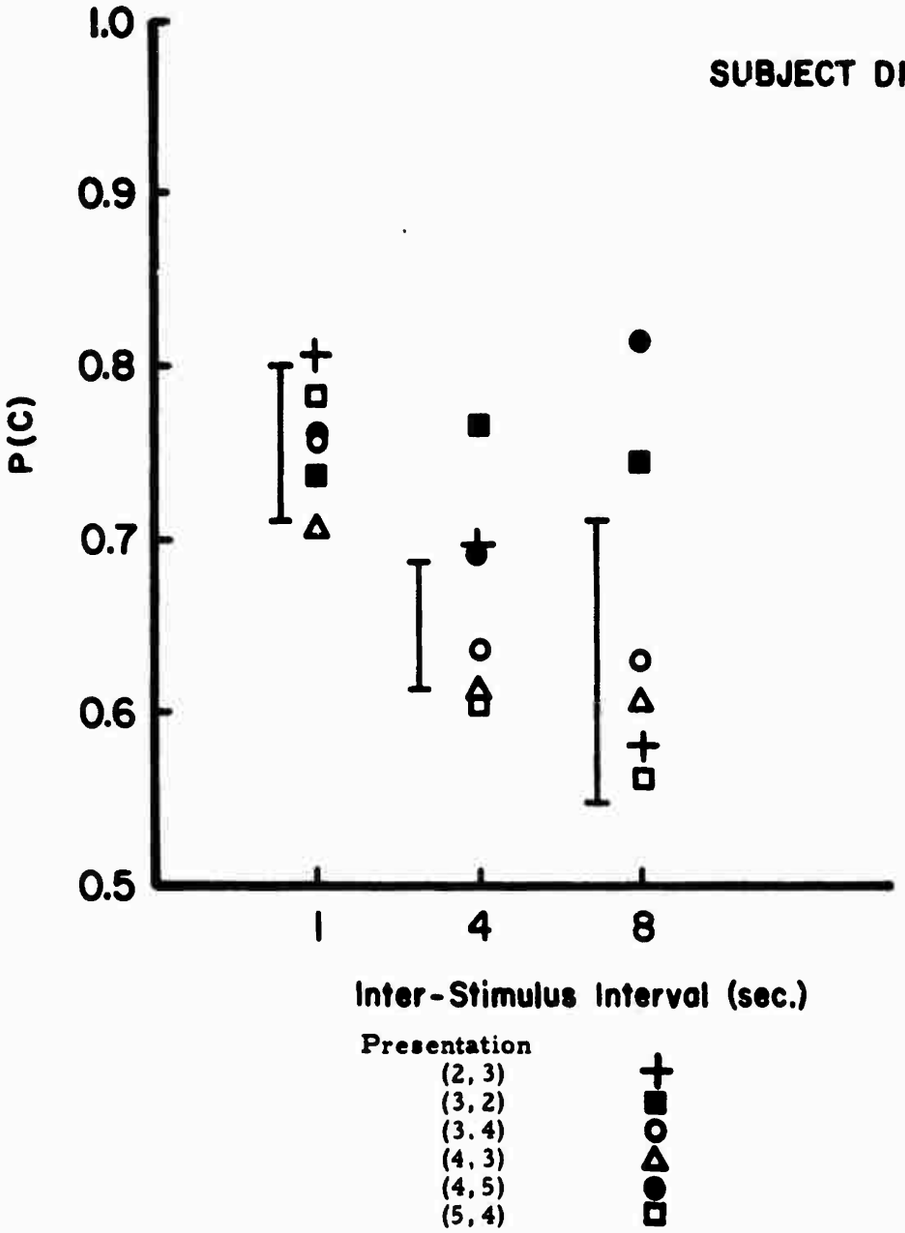


Figure 4: Probability of a correct response from experiment 3 at three inter-stimulus intervals.

from presentations (3, 2) and (4, 5); open points are for the other presentations. Data for each inter-stimulus interval is plotted separately; each point for the 1, 4 and 8 second conditions is based on an average of 250, 300 and 150 observations, respectively. A typical 95% confidence interval, for presentation (3, 4), is included at each inter-stimulus interval.

Discussion of experiment 3.

Figure 4 shows that presentations (3, 2) and (4, 5) did not lead to more correct responses for an inter-stimulus interval of 1 second. At 4 seconds these two presentations begin to show some evidence of being more discriminable, and at 8 seconds the effect appears definite. The discriminability for these two presentations appears to increase as a function of the inter-stimulus interval, while the other four show a steady decline in percent correct responses. It was this effect which was to be the basis for partitioning the behavior that would correct for partial identification responding in the method of constant stimuli.

The logic behind this procedure seemed compelling at the time, particularly since the data of Figure 4 seemed to confirm the notion. It was reasoned that the differences in percent correct responding between the two presentation sets could be used as an estimate of the gain in discriminability which occurred as a result of including partial identification responses in some presentations while excluding them from others. The crucial point was the exclusion of partial identification responses. Unfortunately, more careful thought revealed that although the procedure made a statistical correction for the effect of these responses over a long run of trials, there was no provision to influence the response on any given trial. There was every reason to expect that identification responses were as frequent to the "corrected" presentations as to the others.

The temptation was to assume that on each trial, one or the other response system was used, and that over a block of trials, the observed data could be represented by the addition of two confusion matrices. One matrix, representing trials when actual discriminative responses were made, was hypothesized to be the same under both signal presentation sets. The other matrix was for identification responses, and would reflect the different experimental effects of the two presentation sets. Thus the observed confusion matrix for each presentation set would reflect a constant, underlying discriminative effect compounded by an identification aspect which was differentially affected by the special presentation structure of experiment 3. It was not surprising that the number of unknown parameters required for this partitioning outstripped the number of assumptions which could reasonably be made about them. In the face of such difficulties, the attempts to decompose these data were terminated.

It was also evident at this point that there was no single decision axis available for the generation of confidence responses (Egan, Schulman and Greenberg, 1959). Because the source of such responding was unclear, it was decided to abandon the confidence responses. This is why the data in this report which were collected using confidence-ratings are reported in collapsed, binary form (see footnote 2).

The complete failure of this rather elaborate attempt at experimental decomposition suggested that only a model-bound analysis could rescue the data obtained with the method of constant stimuli. The attack turned toward theoretical models for how identification and discrimination responses could be combined to generate the observed behavior.

Experiment 4

Previous work indicated that at least two response systems would have to be present in the model. These systems and the strategy by which they were combined would have to be represented in sufficient detail so as to include factors of ensemble size and composition, stimulus spacing and inter-stimulus intervals. However, an empirical question remained about the ability of observers to partially identify the particular stimulus ensembles used in these experiments. Quantitative data for a single subject required to recognize signals separated by only 0.4 db. and covering intensity ranges as small as 1.2 db. (the range of experiment 3) were not available. Experiment 4 was performed, using the writer as the subject, to provide partial identification data for such narrowly-spaced ensembles.

Single stimulus presentations were obtained from the apparatus of experiment 3. Levels 2, 3, 4 and 5 were used at the same intensity settings and presented at random with probability 1/4. The response alternatives to each tone were "Loud" or "Soft." Over 200 trials were discarded as showing practice effects. In the first condition, no feedback was given and 50 observations were made at each signal, as shown by the open circles of Figure 5. The ordinate here is the probability of a "Loud" response. In the second condition, the feedback was arbitrarily assigned to be "Loud" for signals 4 and 5 and "Soft" for 2 and 3. About 130 trials were run at each signal under these conditions, shown as closed circles in Figure 5.

Discussion of experiment 4.

It came as no surprise that the data of Figure 5 resembled the sort of psychometric functions ordinarily obtained with the method of constant stimuli,

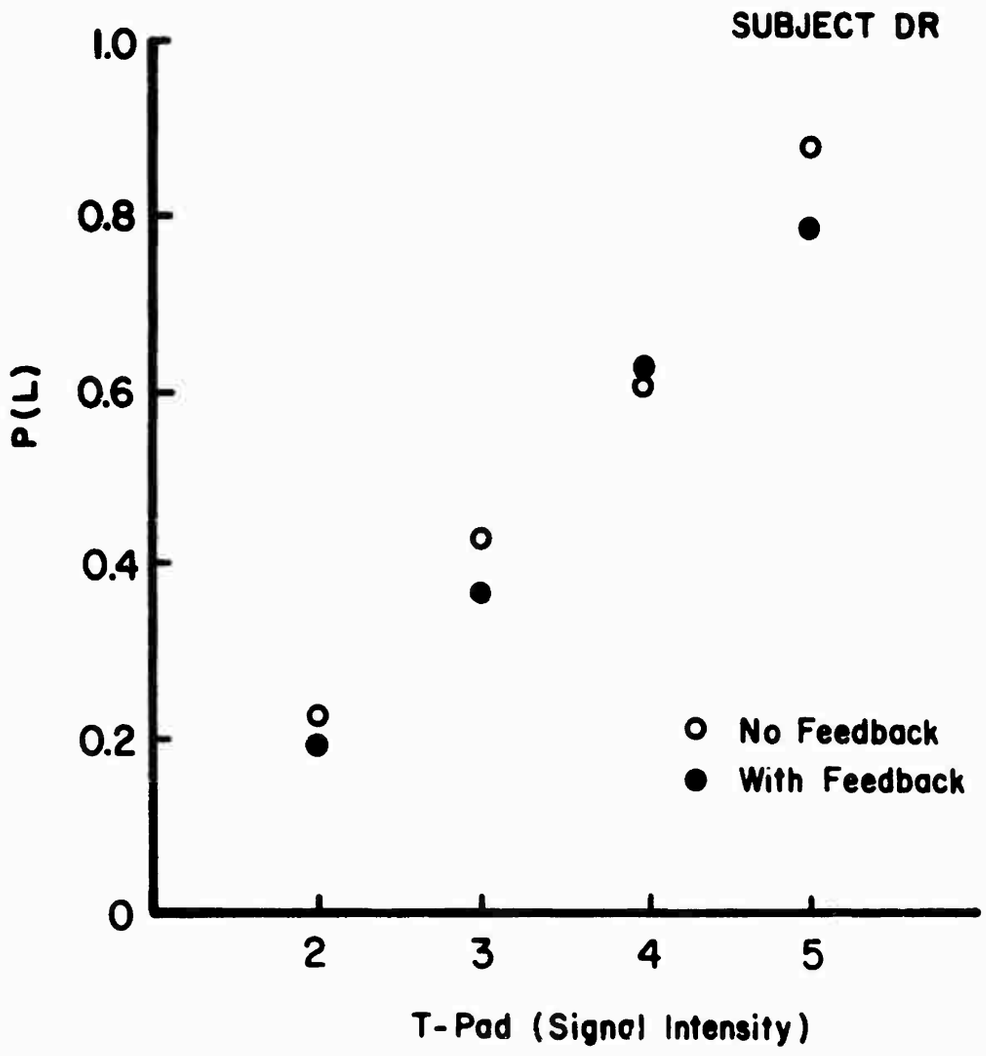


Figure 5: Partial identification data of single stimuli from experiment 4. The ordinate is the probability of a "Loud" response.

for such experiments had been done earlier (Wever and Zener, 1928), although with much greater stimulus ranges. It was revealing, however, to note the relative ease with which these data were produced, both with and without feedback. This result, taken together with the fact that the identification was quite accurate, indicated the importance which such partial identification responses could assume.

A model for quasi-discrimination.

The results of the first four experiments influenced the development of the models for discrimination, which followed Thurstone's notion of a discriminal dispersion (Thurstone, 1959). Repeated presentation of the same signal was assumed to give rise to a distribution of the perceptual effect, Ψ, produced by the signal. Partial identification responses were generated from the Ψ observations by following a cut-point decision rule. Actual discrimination responses were based on differences of successive Ψ's. Parameters separating these distributions were to characterize their identifiability or their discriminability and the variance of the difference distribution was to account for the decline in accuracy of comparison responding as the inter-stimulus interval increased. An additional complication was the set of rules which determined how the two response systems were combined to generate the complex quasi-discrimination behavior.

The most promising model needed so many free parameters that if experiment 3 were used to evaluate goodness-of-fit, there would scarcely be enough degrees of freedom in the data. It was decided that a more adequate test of the model would be to perform two separate experiments. The first would be like experiment 4, and would be used only to provide estimates of parameters, the

separations of the Y distributions. The second experiment would use the same signals, as in experiment 3, at different inter-stimulus intervals. For this experiment the only free parameter, which would have to account for all the degrees of freedom in the data, would be the variance of the difference distribution.

This plan produced a formidable estimation problem. Given a reasonable amount of experimental variability in the parameter estimation experiment, would the parameters be estimated with enough accuracy to verify the model, even if it were true? To provide an answer to this question, a Monte Carlo simulation of the entire procedure was performed on a digital computer. The results of a goodness-of-fit test indicated that if there were no more than binomial variability in the partial identification data and if reasonable sample sizes were taken, the procedure could be successful in validating the model, but not by a large margin.

This model-bound analysis of classical discrimination data was hardly a straightforward approach, especially considering the complexity of the model and the number of free parameters. Misgivings about the whole procedure lead to the abandonment of this approach in favor of another experimental method for obtaining discriminative responses.

An alternative to the classical procedure.

The method of constant stimuli may be designated as the AX method, indicating the first stimulus is always a standard and the second is one of several stimuli. The trouble with the AX method is that is is subject to variables which, putting it loosely, shouldn't matter: context and ensemble effects and identification responses. On the other hand, the procedure is not sensitive to something which ought to matter, the passage of time between stimuli. If an experimental method could be found which was less disturbed by contexts, etc. and more sensitive

to the inter-stimulus interval we would be closer to studying the problem of interest.

The argument for the adoption of another technique might begin by showing how it is logically superior in some sense. The logic of the experimenter would seem to be almost irrelevant in this case however, for there is nothing wrong with the logic of the AX method. It is the behavior of the subject that determines the suitability of procedure in this case, and because of this, the most straightforward test of any alternative methods may be a comparison of real data generated by the competing techniques.

An alternative to the AX technique, which was used for experiments 5 and 6, is a variation of a procedure devised by Munson and Gardner (1950) called the ABX method. In this procedure, three stimuli are presented on every trial, the first two of which are different in some respect, and the third is a repetition of either stimulus A or B. The subject is asked to indicate whether stimulus X was more like A or B. Stimulus X may be physically the same as either A or B, so that the experimenter feels justified in using feedback following responses. By using stimuli which differ on more than one dimension and relaxing the requirement of physical identity, the method can be used for stimuli ordinarily reserved for scaling experiments, omitting feedback in this instance. The ABX method has the feature of presenting on each trial a sample of the signal to be discriminated, that is, the aspect and magnitude in which the stimuli differ are displayed repeatedly.

Because we prefer actual data over logical arguments, we turn now to that evidence. The few times that the method has been used in the literature (Harris, Pikler, Hoffman and Ehmer, 1958; Munson and Gardner, 1950; Munson and Wiener, 1950; Rosenblith and Stevens, 1953) it has been observed that difference limens obtained with the ABX procedure are larger than those from the AX

technique (but see Saslow, 1967). Harris has interpreted this result as showing the superiority of the AX method (Harris, 1952b), but it may also be asserted that this difference is in the desired direction. If a way of responding that is effective under other paradigms is eliminated or reduced by the ABX method, performance ought to be inferior for a given signal strength. As an empirical comparison of the suitability of the ABX and the AX methods, experiment 5 was performed.

Experiment 5

The position of stimulus X in the ABX design is arbitrary, so to make the situation as similar as possible to the classical method, the stimulus to be remembered, the standard, may be presented first, making the form XAB. The temporal intervals during which the three stimuli are presented are called the X interval and comparison intervals 1 and 2, respectively. The number of stimuli and the intensity range which they cover were selected to coincide with those used in experiment 1. For each value of stimulus A and its corresponding B there are four presentations dictated by considerations of symmetry: (AAB), (ABA), (BBA), and (BAB).

The apparatus was basically that of experiment 3, except that changes were made to gate three signals per trial instead of two. To create the four presentations at any given signal level, one of the 8 T-pads was moved to a leading position relative to the other 7. A relay switched this increment pad in or out before every gating of the signal by the electronic switch so that every tone on a given trial had one of two amplitudes. Signal durations were 500 msec. and both inter-stimulus intervals were 1 second. Each trial began with a warning signal, a dim green light that appeared one second before the stimuli. Initially,

three small neon lamps were illuminated successively during the signal presentations as an aid to the subject in keeping track of the particular significance of the tone being heard. It was quickly established that these visual cues were more of a hindrance than an aid to the discrimination, so use of these markers was discontinued early in training. Feedback immediately following responses was given by illumination of the pilot light above the pushbutton labeled 1 or 2, indicating that the tone in the first or second comparison interval contained a repetition of the tone in interval X.

Stimulus sequences on paper tape were in blocks of 110 trials, the first 10 of which were for practice and were not included in the analysis. At the end of each block of trials the experimenter provided feedback regarding the percentage of correct responses via an intercom. The observer was the same one who served in experiments 1 and 2, but experiment 5 was performed over a year later, during which time the subject did not serve in any experiments. Data from 8 sessions, averaging four blocks per session, were discarded as being pre-asymptotic.

In order to maximize the similarity to experiment 1, four signal levels for stimulus A were used that were separated by 5 db., a value intermediate between the 10 db. separation of standards in condition 2 of experiment 1 and the 2 db. separation of condition 4. The four levels were -10, -5, 0 and +5 db. re S_o and the increment was set so that $\Delta = 0.8$ db.

The data were first collected with only one level of signal A occurring on a given day, as in condition 1 of experiment 1. Levels at -10 and 0 db. were used on two separate days to obtain 400 observations at each level. Then all four levels were combined to form an ensemble and more than 500 observations were

collected over several days for each level. Lastly, levels at -5 and +5 db. were used on separate days for 500 trials each. The ensemble data are shown by the open symbols in Figure 6; closed points indicate the data from single level conditions. The abscissa of Figure 6 represents $P[1|X = 2]$, the probability of selecting interval 1, given that repetition of signal X does in fact occur in interval 2. The ordinate is the corresponding "hit" probability, $P[1|X = 1]$.

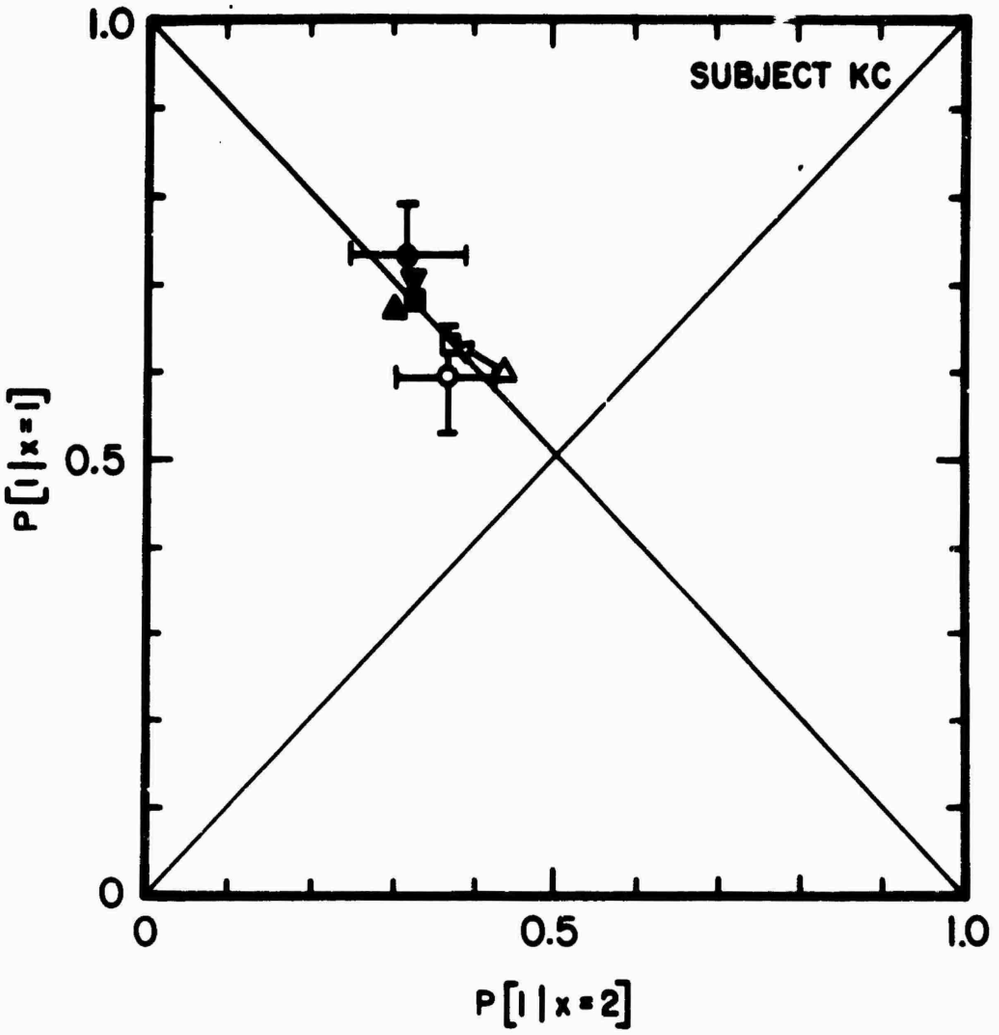
Discussion of experiment 5.

It is evident from Figure 6 that the response bias associated with a particular level is nearly the same, whether the level occurs in an experimental session by itself or within the context of other levels. This result is in contrast to the equivalent data obtained from the AX method as shown in Figure 1. In comparing these two figures it should be noted that the observer is the same in both instances, but that the size of the increment is 0.4 db. in the AX case and 0.8 db. in the XAB instance. For another comparison to AX data, Pollack's graphs may be consulted to show how very much greater the response bias is using the AX method under the same conditions (Pollack, 1956, figure 2, page 908).

Experiment 6 examines the effect upon the XAB method of changing the inter-stimulus interval and the size of the increment. Should the results of these manipulations also favor the XAB procedure, we would have some evidence for preferring it over the AX method.

Experiment 6

All 7 levels of intensity for signal A were used for this experiment, and because each level had four presentations dictated by symmetry considerations, there were 28 different presentations. An intensive evaluation of this procedure was planned in which data from different presentations and levels were to be



Amplitude of signal A in db. re S_0	Single level conditions	Ensemble conditions
+5	■	□
0	●	○
-5	▲	△
-10	▼	▽

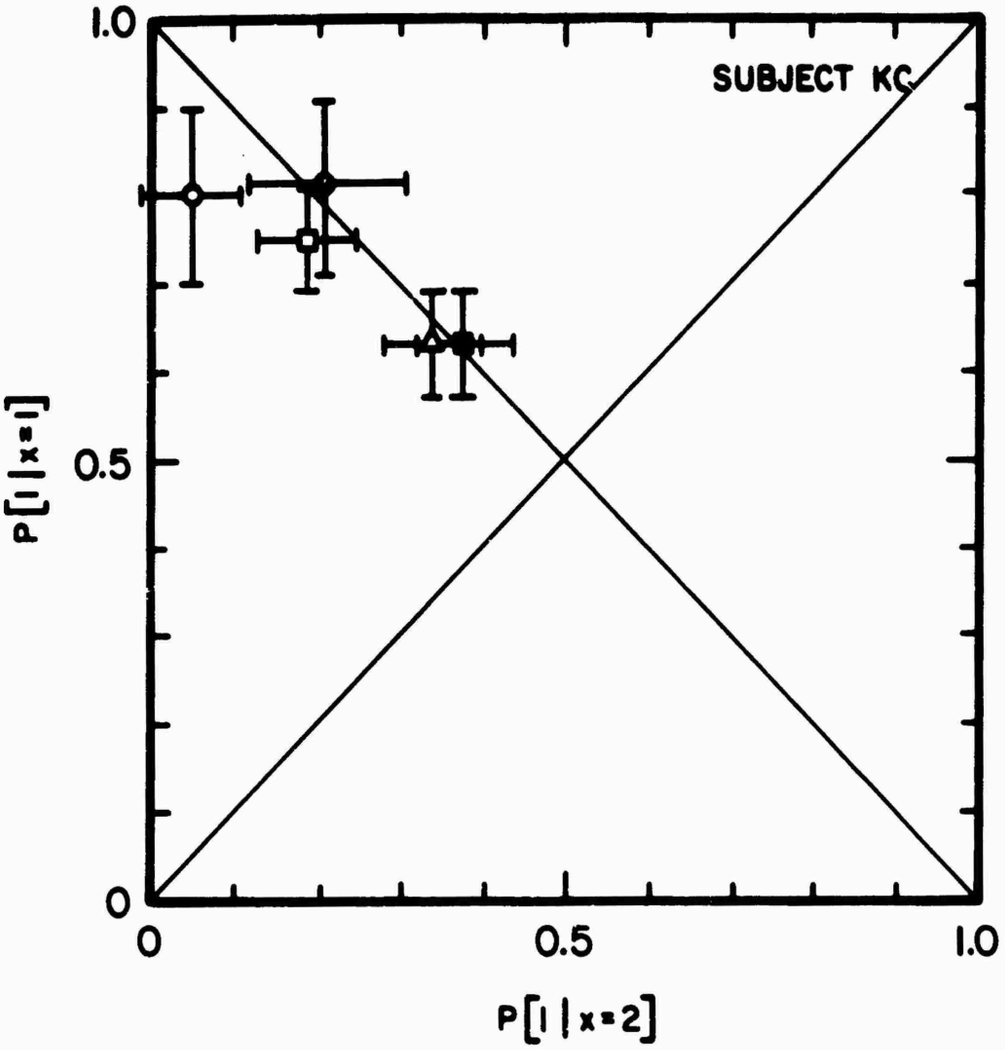
Figure 6: Data from the XAB method under single level and ensemble conditions.

examined in detail. To facilitate these comparisons, it was decided to sample from the set of possible stimulus presentations without replacement. This guaranteed that the random device did not produce deviations from desired presentation probabilities over different blocks of trials as a result of sampling variations. A digital computer performed the randomization process and printed data sheets from which the experimenter entered the information into the apparatus by depressing lever switches on every trial. The computer was programmed in such a way that the sequence never repeated within the number of trials generated, so there was nothing for the subject to learn about a specific stimulus sequence.

There are two inter-stimulus intervals in the XAB procedure, but only the first is varied in this experiment. The duration of the signals was 500 msec. as in experiment 5, but the inter-stimulus intervals were reduced to 500 msec. The 7 levels of signal A were spaced as far apart, 3 db., as constraints of the apparatus and considerations of signal-to-noise ratio would permit. The entire intensity range covered was 18 db.; level 2 was set identical to S_o.

Condition 1.

Five combinations of increment size and inter-stimulus interval were examined using the ensemble of 7 levels. Data for each level indicated that the various levels form a relatively homogeneous set. The only deviation from one level to another is a slight tendency for the more intense levels to be more discriminable. Such small differences were not found to be very informative, so the data for all 7 levels was pooled to form a single datum for each combination of increment and inter-stimulus interval. These data were collected after experiment 5 from the same subject. Figure 7 shows the effect of variations in Δ and the interval on the pooled data, together with the number of observations taken under each condition.



Δ (db.)	ISI (sec.)	N	
4.0	1/2	140	○
4.0	8	140	●
2.5	1/2	560	□
2.5	8	450	■
1.5	1/2	560	△

Figure 7: Data from the XAB method under ensemble conditions at various inter-stimulus intervals and signal strengths.

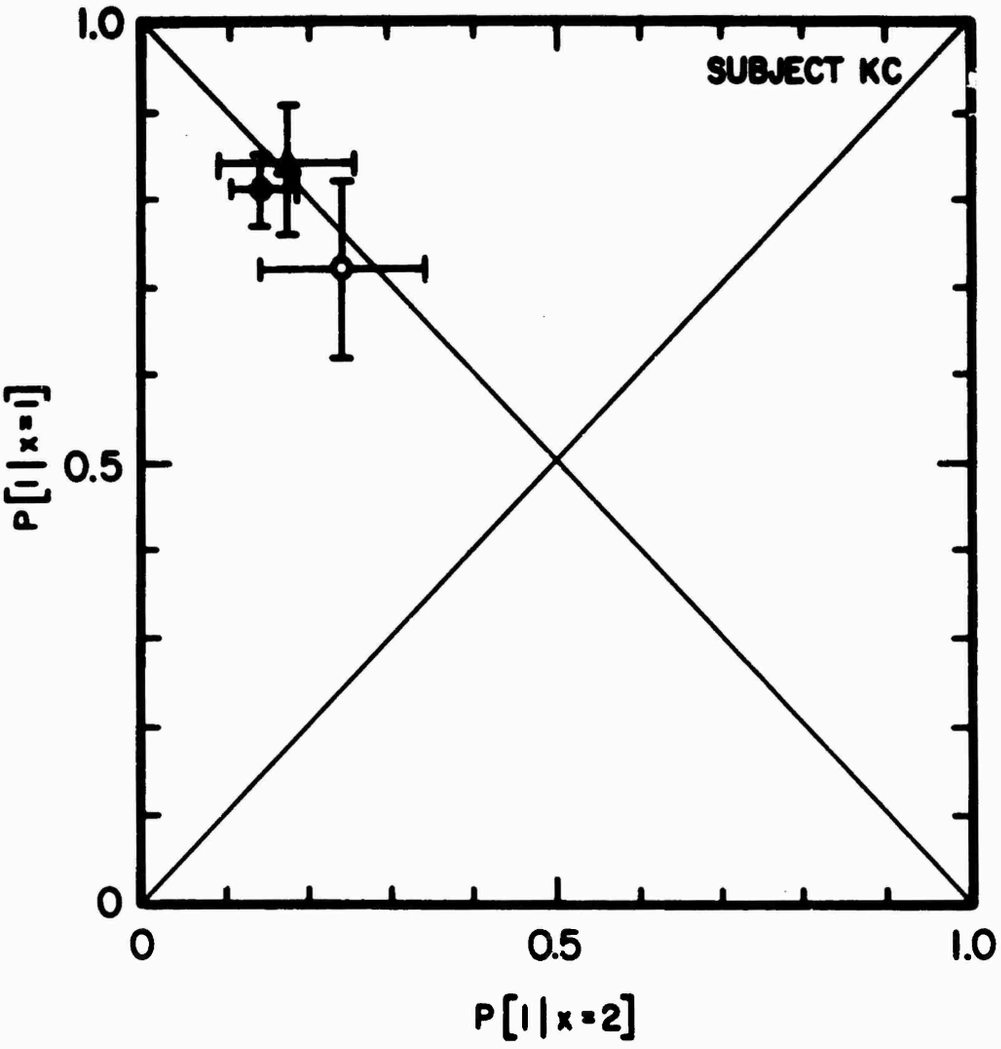
Condition 2.

To demonstrate that there is not something special about the XAB design, that ensemble effects are also present in this method, level 3 was used as the only level for several sessions. This performance, represented by the solid circle in Figure 8, contrasts with the discrimination of level 3 when it occurs in the context of the 7 level ensemble, shown by the open circle (data from condition 1). Finally, the inter-stimulus interval was increased to 8 seconds and level 3 was used alone to obtain the point plotted as a triangle in Figure 8.

Discussion of experiment 6.

Figure 7 demonstrates that when the inter-stimulus interval is fixed, the effective signal strength varies regularly with Δ. Also, when Δ is held fixed at 4.0 db., increasing the interval from 1/2 to 8 seconds decreases the discriminability, although the decrease is larger when Δ = 2.5 db. The large increments required in the XAB procedure using 7 levels were initially of some concern. Fortunately, some relevant AX data on individual, experienced observers discriminating at several levels are available (Pollack, 1956). In this experiment, 100 msec. bursts of 1000 Hz were presented with an inter-stimulus interval of 1.2 seconds. Ten levels were equally spaced in decibels, the spacing varied from 0 to 8 db. and the levels were presented in random order. Pollack's finding was that the DL increases as the range of variation of the levels increases. The case where the levels are 2 db. apart provides a total range of 18 db. and matches the range of experiment 6. Inspection of Pollack's graphs shows that under such conditions the DL for 70% correct discrimination varies between about 2 and 3 db., in close agreement with the data of Figure 7.

The change in effective signal strength which results from a change in ensemble size is evident in Figure 8. The fact that the inter-stimulus interval



	Δ(db.)	ISI (sec.)	N	
Level 3 singly	2.5	1/2	720	●
Level 3 singly	2.5	8	200	▲
Level 3 in ensemble	2.5	1/2	180	○

Figure 8: Data from the XAB method contrasting ensemble with single level effects.

has little effect (as in the AX method, Figure 2) when only a single level is used also demonstrates the importance of the ensemble for the XAB procedure.

Conclusion

When experienced observers are used and data from single subjects is examined, it is not easy to obtain data which meet our intuitive requirements about what should constitute discrimination behavior. It is seen that observers can utilize cues which are not purposely made relevant by the experimenter when the classical AX procedure is used. These para-discriminative cues seem to play an important role in determining response bias under ensemble conditions and in aiding discriminability when long inter-stimulus intervals are used.

A method has been proposed and examined that reduces the undesirable aspects of the classical procedure and which may be useful in critical applications. Experimentation is currently under way using this procedure which will provide more extensive discrimination data.

References

Bressler, J. Judgments in absolute units as a psychophysical method. Arch. Psychol., 1933, 23, No. 152, 1-68.

Bush, R.R., Galanter, E. and Luce, R.D. Characterization and classification of choice experiments. In R.D. Luce, R.R. Bush and E. Galanter (Eds.) Handbook of Mathematical Psychology, Vol. 1. New York: Wiley, 1963.

Doughty, J.M. The effect of psychophysical method and context on pitch and loudness functions. J. exp. Psychol., 1949, 39, 729-745.

Egan, J.P., Schulman, A.I. and Greenberg, G. Operating characteristics determined by binary decisions and by ratings. J. acoust. Soc. Amer., 1959, 31, 768-773.

Eriksen, C.W. and Hake, H.W. Absolute judgments as a function of stimulus range and number of stimulus and response categories. J. exp. Psychol., 1955, 49, 323-332.

Galanter, E. Contemporary Psychophysics. In New Directions in Psychology. New York: Holt, Rinehart and Winston, 1962.

Garner, W.R. An informational analysis of absolute judgments of loudness. J. exp. Psychol., 1953, 46, 373-380.

Garner, W.R. and Hake, H.W. The amount of information in absolute judgments. Psychol. Rev., 1951, 58, 446-459.

Gourevitch, V. and Galanter, E. A significance test for one parameter isosensitivity functions. Psychophysics Laboratory Report No. PLR-19A, 1966, University of Washington.

Green, D.M. and Swets, J.A. Signal Detection Theory and Psychophysics. New York: Wiley, 1966.

Harris, J.D. Discrimination of pitch: suggestions toward method and procedure. Amer. J. Psychol., 1948, 61, 309-322.

Harris, J.D. The effect of inter-stimulus interval on intensity discrimination for white noise. Amer. J. Psychol., 1949, 62, 202-214.

Harris, J.D. The decline of pitch discrimination with time. J. exp. Psychol., 1952a, 43, 96-99.

Harris, J.D. Remarks on the determination of a differential threshold by the so-called ABX technique. J. acoust. Soc. Amer., 1952b, 24, 417.

Harris, J.D., Pikler, A.G., Hoffman, H.S. and Ehmer, R.H. The interaction of pitch and loudness discriminations. J. exp. Psychol., 1958, 56, 232-238.

Koester, T. and Schoenfeld, W.N. The effect of context upon judgments of pitch differences. J. exp. Psychol., 1946, 36, 417-430.

König, E. Effect of time on pitch discrimination thresholds under several psychophysical procedures; comparison with intensity discrimination thresholds. J. acoust. Soc. Amer., 1957, 29, 606-612.

Luce, R.D. Detection and recognition. In R.D. Luce, R.R. Bush and E. Galanter (Eds.) Handbook of Mathematical Psychology, Vol 1. New York: Wiley, 1963.

Luce, R.D. and Galanter, E. Discrimination. In R.D. Luce, R.R. Bush and E. Galanter (Eds.) Handbook of Mathematical Psychology, Vol. 1. New York: Wiley, 1963.

Munson, W.A. and Gardner, M.B. Loudness patterns - a new approach. J. acoust. Soc. Amer., 1950, 22, 177-190.

Munson, W.A. and Wiener, F.M. Sound measurements for psychophysical tests. J. acoust. Soc. Amer, 1950, 22, 382-386.

Needham, J. G. The time-error in comparison judgments. Psychol. Bull., 1934(a), 31, 229-243.

Needham, J. G. The time-error as a function of continued experimentation. Amer. J. Psychol., 1934(b), 46, 558-567.

Needham, J. G. The effect of the time interval upon the time-error at different intensive levels. J. exp. Psychol., 1935, 18, 530-543.

Pollack, I. Intensity discrimination thresholds under several psychophysical procedures. J. acoust. Soc. Amer., 1954, 26, 1056-1059.

Pollack, I. Identification and discrimination of components of elementary auditory displays. J. acoust. Soc. Amer., 1956, 28, 906-909.

Postman, L. The time-error in auditory perception. Amer. J. Psychol., 1946, 59, 193-219.

Postman, L. Time-error as a function of the method of experimentation. Amer. J. Psychol., 1947, 60, 101-108.

Ronken, D. A. and Galanter, E. Studies of the constant error. Psychophysics Laboratory Report, PLR-14N, 1965, University of Washington.

Rosenblith, W. A. and Stevens, K. W. On the DL for frequency. J. acoust. Soc. Amer., 1953, 25, 980-985.

Saslow, M. G. Frequency discrimination as measured by AB and ABX procedures. J. acoust. Soc. Amer., 1967, 41, 220-221.

Saslow, M. G. and Markowitz, H. An inexpensive and simple solid state timer. J. exp. Anal. Behav., 1964, 7, 252.

Tanner, W. P. and Rivette, C. L. Learning in psychophysical experiments. J. acoust. Soc. Amer., 1963, 35, 1896. (Abstract).

Thurstone, L.L. The Measurement of Values. Chicago: University of Chicago Press, 1959.

Wever, E.G. and Zener, K.E. The method of absolute judgment in psychophysics. Psychol. Rev., 1928, 35, 466-493.

Whipple, G.M. An analytic study of the memory image and the process of judgment in the discrimination of clangs and tones. Amer. J. Psychol., 1901, 4, 409-457.

Woodrow, H. Weight discrimination with a varying standard. Amer. J. Psychol., 1933, 45, 391-416.

Woodworth, R.D. Experimental Psychology. New York: Holt, 1938.

Wright, H.N. Audibility of switching transients. J. acoust. Soc. Amer., 1960, 32, 138.

Footnotes

[1] Portions of this research were supported by Contract Nonr 477(34) between the Office of Naval Research and the University of Washington. This work was conducted during the author's pre-doctoral traineeship provided by the National Aeronautics and Space Administration.

[2] There were in fact 8 response alternatives available to the subject, because the collection of these data included confidence-rating responses. The binary, louder-softer partitioning was specified as the first aspect of the response. The subject then subdivided this into four levels of confidence. As will be shown, the results of this and subsequent experiments have demonstrated complexities which leave the source of the rating response in doubt. For this reason, only the collapsed, binary form of the response is reported here. Because the binary judgment was in fact made by the subject, it is permissible to make what would otherwise be a rather arbitrary collapsing of the rating data.

[3] The next two experiments will show that this test is not strictly applicable because the assumptions used in its derivation are violated by the data. Nevertheless, the test results are presented as the most appropriate quantitative measure available at this time.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* University of Washington Seattle, Washington 98105 Eugene Galanter, Principal Investigator	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

AN EXPERIMENTAL CRITIQUE OF THE METHOD OF CONSTANT STIMULI AND SOME ALTERNATIVE PROCEDURES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

A technical report.

5. AUTHOR(S) *(First name, middle initial, last name)*

Don A. Ronken

6. REPORT DATE 15 April 1967	7a. TOTAL NO. OF PAGES 33	7b. NO. OF REFS 39
8a. CONTRACT OR GRANT NO. NONR 477(34) b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) PRP-31N 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research

13. ABSTRACT

Some experiments investigating the constant error demonstrate that the method of constant stimuli is especially unsuited for studying such small effects of discrimination. The nature of the confounded data from the classical procedure suggests that the difficulty is of a fundamental nature and can be expected to influence the data under less stringent conditions as well. Some data from the literature are offered to support this view. Several alternative procedures are developed and evaluated experimentally, the most promising of which is shown to be a form of the XAB method.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT

DD FORM 1 NOV 68 1473 (BACK)
(PAGE 2)